A maximum likelihood method for bidimensional experimental distributions, and its application to the galaxy merger fraction

Carlos López-Sanjuan, César Enrique García-Dabó, Marc Balcells
Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, La Laguna, Tenerife, 38200 Spain
clsj@iac.es, enrique.garcia@gtc.iac.es, balcells@iac.es

ABSTRACT

The determination of galaxy merger fraction of field galaxies using automatic morphological indices and photometric redshifts is affected by several biases if observational errors are not properly treated. Here, we correct these biases using maximum likelihood techniques. The method takes into account the observational errors to statistically recover the real shape of the bidimensional distribution of galaxies in redshift - asymmetry space, needed to infer the redshift evolution of galaxy merger fraction. We test the method with synthetic catalogs and show its applicability limits. The accuracy of the method depends on catalog characteristics such as the number of sources or the experimental error sizes. We show that the maximum likelihood method recovers the real distribution of galaxies in redshift and asymmetry space even when binning is such that bin sizes approach the size of the observational errors. We provide a step-by-step guide to applying maximum likelihood techniques to recover any one- or bidimensional distribution subject to observational errors.

Subject headings: Data Analysis and Techniques

1. INTRODUCTION

The currently popular hierarchical ΛCDM models are successful at explaining the structure build-up of the cold dark matter component of the Universe (Springel et al. 2005). But such models have difficulties when explaining the evolution of the baryonic component, even with modeling that incorporates star formation, active galactic nuclei and supernova feedback, and the multiphase nature of the interstellar medium (De Lucia & Blaizot 2007, and references therein). An open question is the role of the galaxy mergers in the formation of today's galaxies, specially the most massive ellipticals. The observational determination of the merger rate, $\Re_{\rm m}$, and its evolution with redshift, provide empirical clues on the amount and the timing of the merger activity. They also constitute key inputs for semi-analytic models of galaxy formation and evolution.

The merger rate, defined as the number density of merger systems at given redshift, depends on the merger time τ_m, which can only be estimated by N-body simulations and simplified models (Mihos 1995; Patton et al. 2000; Conselice 2006). On the other hand, the galaxy merger fraction $f_{\rm gm}$, defined as the number of merger galaxies in a given galaxy sample in a redshift interval, is a direct observational quantity. Many works have determined the galaxy merger fraction, usually parametrized as $f_{\rm gm} = f_{\rm gm,0} \cdot (1+z)^m$, using different sample selection and methods, like morphological criteria (Conselice 2003; Lavery et al. 2004; Cassata et al. 2005; Lotz et al. 2008; Bridge et al. 2007; De Propris et al. 2007), kinematic close companions (Patton et al. 2000, 2002; Lin et al. 2004; De Propris et al. 2005, 2007), spatial close pairs (Le Fèvre et al. 2000; Bundy et al. 2004; Bridge et al. 2007; Kartaltepe et al. 2007) or correlation function (Bell et al. 2006; Masjedi et al. 2006). In these works the value of the merger index varies in the range $m = 0-4$. ΛCDM models predict $m \sim 3$ (Kolatt et al. 1999; Governato et al. 1999; Gottlöber et al. 2001).

The morphological criterion for determining the galaxy merger fraction (see Conselice 2003, hereafter C03), is based on the fact that, just after a merger is complete, the galaxy image shows strong geometrical distortions, in particular asymmetric distortions. Hence, high values in the automatic asymmetry index A (Abraham et al. 1996; C03) are assumed to identify merger systems. Other automatic morphological indices, such as M_{20} and G, have also been used to determine the evolution of galaxy merger fraction with redshift (Lotz et al. 2008). The determination of morphological indices, which must be done on HST images, is affected by surface brightness dimming and K-corrections, so the errors of the indices grow with redshift and are more important for faint galaxies.

In this paper, we present a method based on the maximum likelihood (ML) technique, to handle the effects of the large errors on the determination of the galaxy merger fraction. Galaxy Merger fraction determinations using morphological criteria are generally done on large photometric surveys such as AEGIS (Davis 2007), COMBO-17 (Wolf et al. 2003), COSMOS (Scoville et al. 2007), GOODS (Giavalisco et al. 2004), or SWIRE (Lonsdale et al. 2003). We therefore address the effects of errors in the galaxy asymmetry indices as well as errors on the photometric redshifts.

In Section 2 we review the maximum likelihood method for determining bidimensional distributions. Its application to the galaxy merger fraction determination is given in Section 2.2. These sections have a high mathematical content, and a statistics background is recommended. Then, in Section 3 we summarize the simulations made to test the general method and how it improves the galaxy merger fraction determination, Section 3.7. In Section 4 we provide an outline for the application of the ML method to any one- or bidimensional experimental distribution subject to observational errors. Our conclusions are presented in Section 5.

2. METHODOLOGY

Following Conselice (2006), we define the galaxy merger fraction by morphological criteria as

$$f_{\rm gm} = \frac{\kappa \cdot N_{\rm m}}{N_{\rm tot} + (\kappa - 1) N_{\rm m}}, \qquad (1)$$

where $N_{\rm m}$ is the number of the distorted sources in the sample, classified as the systems with a value in the asymmetry index A higher than a limiting value $A_{\rm m}$ (see C03 for details), $N_{\rm tot}$ is the total number of sources in the sample, and κ is the average number of galaxies that merged to produce the $N_{\rm m}$ merger systems. We use $\kappa = 2$ throughout this paper.

In order to compute the galaxy merger fraction and its redshift evolution we must know the underlying distribution of the z and A values, that we assume is represented by a bidimensional histogram in redshift and asymmetry space. This bidimensional histogram is defined by the number of sources in each redshift-asymmetry bin. Normalizing to unity the histogram yields a bidimensional probability distribution defined now by p_{kl}, the probability that a source has redshift in bin k and asymmetry in bin l. Index k scans the redshift bins of size Δz and index l scans the asymmetry bins of size ΔA. In our case we just need two asymmetry bins separated by $A_{\rm m}$: the $l = 0$ bin represents normal sources and the $l = 1$ bin represents merger systems. Now, the galaxy merger fraction in redshift bin $[z_k, z_{k+1})$ is

$$f_{{\rm gm},k} = \frac{2p_{k1}}{p_{k0} + 2p_{k1}}. \qquad (2)$$

The accuracy with which the p_{kl} can be obtained degrades significantly when photometric redshifts, $z_{\rm phot}$, are used, and for typical errors of A in deep HST surveys. This introduces strong biases in the determination of the galaxy merger fraction.

2.1. The maximum likelihood method

The maximum likelihood method (ML method) developed here is based on García-Dabó (2002), who used this technique to determine unbiased luminosity functions. ML methods have been used in a wide range of topics in astrophysics. Arzner et al. (2007) use it to improve the determination of faint X-ray spectra; Sheth (2007) to obtain redshift and luminosity distributions in photometric surveys; Naylor & Jeffries (2006) to fit colour-magnitude diagrams; Makarov et al. (2006) to improve distance estimates using Red Giant Branch stars; and, Efstathiou (2004) to analyze low cosmic microwave background multipoles from the Wilkinson Microwave Anisotropy Probe. ML

methods are based on the estimation of the most probable values of a set of parameters which define the probability distribution that describes an observational sample (Davidson & Mackinnon 1993; Peña 2001).

The general ML method operates as follows. Throughout the paper we denote as $P(\mathbf{a}|\mathbf{b})$ the probability to obtain the values $\mathbf{a}$, given parameters $\mathbf{b}$. Being $\mathbf{x}_i$ a vector containing all the measured values for source i in the data set and θ the parameters of the underlying multidimensional distribution that we want to estimate, we may express the joined likelihood function as

$$L(\mathbf{x}_i|\theta) \equiv -\ln\big[\prod_i P(\mathbf{x}_i|\theta)\big] = -\sum_i \ln\big[P(\mathbf{x}_i|\theta)\big], \tag{3}$$

where $P(\mathbf{x}_i|\theta)$ is the probability to obtain $\mathbf{x}_i$ for a given θ. If we are able to express $P(\mathbf{x}_i|\theta)$ analytically, we can minimize Equation 3 to obtain the best estimation of parameters θ, denote as θ_{ML}. In our case, $\mathbf{x}_i$ are the observed values of z and A for source i, $\mathbf{x}_i \equiv (z_{\mathrm{obs},i}, A_{\mathrm{obs},i})$, while $\theta \equiv (p_{kl}, \alpha)$, where p_{kl} are the probabilities which we defined in the paragraph previous to Equation 2, and α denotes any other fixed parameters of the distribution.

Sources are assumed to have real redshift and asymmetry values $z_{\mathrm{real},i}$ and $A_{\mathrm{real},i}$ (not affected by observational errors) which define a bidimensional distribution p_{kl} such that

$$P_{2D}(z_{\mathrm{real},i}, A_{\mathrm{real},i}|p_{kl}) = \{p_{kl}, \forall z_k \leq z_{\mathrm{real},i} < z_{k+1}, A_l \leq A_{\mathrm{real},i} < A_{l+1}\}. \tag{4}$$

Observational errors cause the observed $z_{\mathrm{obs},i}$ and $A_{\mathrm{obs},i}$ to differ from their respective real values $z_{\mathrm{real},i}$ and $A_{\mathrm{real},i}$. The observed $z_{\mathrm{obs},i}$ are assumed to be extracted for a Gaussian distribution with mean $z_{\mathrm{real},i}$ and standard deviation $\sigma_{z_{\mathrm{obs},i}}$,

$$P_G(z_{\mathrm{obs},i}|z_{\mathrm{real},i}, \sigma_{z_{\mathrm{obs},i}}) = \frac{1}{\sqrt{2\pi}\sigma_{z_i}} \mathrm{e}^{-\frac{(z_{\mathrm{obs},i}-z_{\mathrm{real},i})^2}{2\sigma^2_{z_{\mathrm{obs},i}}}}. \tag{5}$$

Similarly, the observed asymmetry values $A_{\mathrm{obs},i}$ are assumed to be extracted from a Gaussian distribution with mean $A_{\mathrm{real},i}$ and standard deviation $\sigma_{A_{\mathrm{obs},i}}$,

$$P_G(A_{\mathrm{obs},i}|A_{\mathrm{real},i}, \sigma_{A_{\mathrm{obs},i}}) = \frac{1}{\sqrt{2\pi}\sigma_{A_{\mathrm{obs},i}}} \mathrm{e}^{-\frac{(A_{\mathrm{obs},i}-A_{\mathrm{real},i})^2}{2\sigma^2_{A_{\mathrm{obs},i}}}}. \tag{6}$$

While the z_{phot} errors may not be strictly Gaussian, this is the best analytical approximation of the errors that we can make. We obtain the probability $P(\mathbf{x}_i|\theta)$ of each source by the total probability theorem:

$$\begin{aligned} &P(z_{\mathrm{obs},i}, A_{\mathrm{obs},i}|p_{kl}, \sigma_{z_{\mathrm{obs},i}}, \sigma_{A_{\mathrm{obs},i}}) \\ &= \int P_G(z_{\mathrm{obs},i}|z_{\mathrm{real},i}, \sigma_{z_{\mathrm{obs},i}}) \\ &\times P_G(A_{\mathrm{obs},i}|A_{\mathrm{real},i}, \sigma_{A_{\mathrm{obs},i}}) \\ &\times P_{2D}(z_{\mathrm{real},i}, A_{\mathrm{real},i}|p_{kl})\mathrm{d}z_{\mathrm{real},i}\mathrm{d}A_{\mathrm{real},i}, \end{aligned} \tag{7}$$

where $\mathbf{x}_i \equiv (z_{\mathrm{obs},i}, A_{\mathrm{obs},i})$ and $\theta \equiv (p_{kl},\ \sigma_{z_{\mathrm{obs},i}},\ \sigma_{A_{\mathrm{obs},i}})$ in Equation 3, with $\alpha \equiv (\sigma_{z_{\mathrm{obs},i}}, \sigma_{A_{\mathrm{obs},i}})$. Note that the values of $\sigma_{z_{\mathrm{obs},i}}$ and $\sigma_{A_{\mathrm{obs},i}}$ are the measured uncertainties for each source, so the only unknowns are the probabilities p_{kl}, which we want to estimate. Note also that we integrate over the variables $z_{\mathrm{real},i}$ and $A_{\mathrm{real},i}$, so we are not be able to estimate them individually, but only the underlying bidimensional distribution p_{kl} that describes the sample.

In order to ensure that the probabilities p_{kl} are not negative, we change variables, $p_{kl} = \exp(p'_{kl})$; this change keeps our problem analytic. With these new variables and after integrating Equation 7, our likelihood function, defined in Equation 3, becomes

$$L(z_{\mathrm{obs},i}, A_{\mathrm{obs},i}|p'_{kl}, \sigma_{z_{\mathrm{obs},i}}, \sigma_{A_{\mathrm{obs},i}}) = \sum_i \left[\ln\left\{\sum_k \sum_l \frac{\mathrm{e}^{p'_{kl}}}{4}\mathrm{ERF}(z,i,k)\mathrm{ERF}(A,i,l)\right\}\right], \tag{8}$$

where

$$\mathrm{ERF}(\eta,i,k) = \mathrm{erf}\left(\frac{\eta_{\mathrm{obs},i}-\eta_{k+1}}{\sqrt{2}\sigma_{\eta_{\mathrm{obs},i}}}\right) - \mathrm{erf}\left(\frac{\eta_{\mathrm{obs},i}-\eta_k}{\sqrt{2}\sigma_{\eta_{\mathrm{obs},i}}}\right), \tag{9}$$

and $\mathrm{erf}(x)$ is the error function. We must observe that in the minimization of Equation 8 the variables p'_{kl} are not independent. This is due to the

normalization of the distribition: the integration over all parameters space muts be one. This impose the following condition over p'_{kl}:

$$\mathbf{g}(p'_{kl}) \equiv \sum_k \sum_l \mathrm{e}^{p'_{kl}} (z_{k+1} - z_k)(A_{l+1} - A_l) - 1 = 0. \tag{10}$$

The method for finding the extrema of a function of several variables subject to one or more constraints is know as the Lagrange multipliers (see e.g., Marsden & Tromba 1996, for details). It states that the function to minimize is not the target function, Equation 8, but a related one:

$$G(p'_{kl}, \lambda) = L(z_{\mathrm{obs},i}, A_{\mathrm{obs},i} | p'_{kl}, \sigma_{z_{\mathrm{obs},i}}, \sigma_{A_{\mathrm{obs},i}}) + \lambda \mathbf{g}(p'_{kl}), \tag{11}$$

where λ is an auxiliary variable. Minimizing Equation 11 we obtain the best p'_{kl} values, denoted as $p'_{kl,\mathrm{ML}}$.

The minimization of Equation 11 can be performed with any numerical minimization code. We used `AMOEBA`, which is based on the commonly used algorithm of Nelder-Mead (Nelder & Mead 1965) and coded in C (Press 1995, pp. 408-412).

At this point we have the probabilities $p'_{kl,\mathrm{ML}}$. However, our goal is to obtain not only the best probabilities estimation, but also their associated uncertainties. The ML method states that we can obtain all the parameter covariaces using an expansion of the function $G(p'_{kl}, \lambda)$ in Taylor's series of its variables $\theta = (p'_{kl}, \lambda)$ around the minimization point $\theta_{ML} = (p'_{kl,\mathrm{ML}}, \lambda_{ML})$ if the probability distributions of $p'_{kl,\mathrm{ML}}$ are Gaussian, which we assume. The previous minimization process made the first G derivate null at $\theta = \theta_{ML}$ and we obtain

$$G = G(\theta_{ML}) + \frac{1}{2}(\theta - \theta_{ML})^T H (\theta - \theta_{ML}), \tag{12}$$

where $H = h_{xy}$ is the Hessian matrix and T denotes the transpose vector. The inverse of the Hessian matrix gives us an estimate of the 68% confidence intervals of $p'_{kl,\mathrm{ML}}$, denoted as $[p'_{kl,\mathrm{ML}} - \sigma_{p'_{kl,\mathrm{ML}}}, p'_{kl,\mathrm{ML}} - \sigma_{p'_{kl,\mathrm{ML}}}]$, and the covariances between each $p'_{kl,\mathrm{ML}}$, denoted as $\mathrm{cov}(p'_{mn,\mathrm{ML}}, p'_{st,\mathrm{ML}})$, because maximum likelihood theory states that $\mathrm{cov}(\theta_x, \theta_y) \geq h^{-1}_{xy}$ and $\sigma_{\theta_x} \geq h^{-1}_{xx}$. In our case, the Hessian matrix is

$$H = \begin{pmatrix} \frac{\partial^2 G}{\partial p'_{mn} \partial p'_{st}} & \nabla g \\ \nabla g & 0 \end{pmatrix}, \tag{13}$$

where

$$\frac{\partial^2 G}{\partial p'_{mn} \partial p'_{st}} = -\sum_i \frac{\mathrm{ERF}(z, i, m)\mathrm{ERF}(A, i, n)}{16} \times \frac{\mathrm{ERF}(z, i, s)\mathrm{ERF}(A, i, t)\mathrm{e}^{p'_{mn}}\mathrm{e}^{p'_{st}}}{\sum_l \sum_k \frac{\mathrm{e}^{p'_{kl}}}{4}\mathrm{ERF}(z, i, k)\mathrm{ERF}(A, i, l)} \tag{14}$$

$$\nabla g = \frac{\partial^2 G}{\partial \lambda \partial p'_{mn}} = (z_{m+1} - z_m)(A_{n+1} - A_n)\mathrm{e}^{p'_{mn}}. \tag{15}$$

Finally, the $p_{kl,\mathrm{ML}}$ probabilities simply are:

$$p_{kl,\mathrm{ML}} = \mathrm{e}^{p'_{kl,\mathrm{ML}}}. \tag{16}$$

Assuming that the $p'_{kl,\mathrm{ML}}$ follow a Gaussian distribution, which is assured by the ML theory for large number of sources, the $p_{kl,\mathrm{ML}}$ follow a log-normal distribution:

$$P_{LN}(p_{kl} | p'_{kl,\mathrm{ML}}, \sigma_{p'_{kl,\mathrm{ML}}}) = \frac{\mathrm{e}^{-(\ln p_{kl} - p'_{kl,\mathrm{ML}})^2 / 2\sigma^2_{p'_{kl,\mathrm{ML}}}}}{\sqrt{2\pi} p_{kl} \cdot \sigma_{p'_{kl,\mathrm{ML}}}}, \tag{17}$$

which is highly asymmetric and whose 68% confidence interval is $[\sigma^-_{p_{kl,\mathrm{ML}}}, \sigma^+_{p_{kl,\mathrm{ML}}}]$, where

$$\sigma^-_{p_{kl,\mathrm{ML}}} = \mathrm{e}^{-\sigma_{p'_{kl,\mathrm{ML}}}} p_{kl,\mathrm{ML}}, \tag{18}$$

$$\sigma^+_{p_{kl,\mathrm{ML}}} = \mathrm{e}^{\sigma_{p'_{kl,\mathrm{ML}}}} p_{kl,\mathrm{ML}}. \tag{19}$$

Furthermore, each p'_{k0} and p'_{k1} are connected by the covariance $\mathrm{cov}(p'_{k0,\mathrm{ML}}, p'_{k1,\mathrm{ML}})$, so the confidence intervals of p_{k0} and p_{k1} are not independent. In the next section we explain how to obtain the confidence interval of the galaxy merger fraction taking this into account.

2.2. The galaxy merger fraction

Expressing the galaxy merger fraction in the range $[z_k, z_{k+1})$ (Equation 1) as a function of the output variables of the ML method we obtain:

$$f^{\mathrm{ML}}_{\mathrm{gm},k} = \frac{2p_{k1,\mathrm{ML}}}{p_{k0,\mathrm{ML}} + 2p_{k1,\mathrm{ML}}}. \tag{20}$$

However, we cannot obtain the 68% confidence interval of $f^{\mathrm{ML}}_{\mathrm{gm},k}$, defined as $[\sigma^-_{f^{\mathrm{ML}}_{\mathrm{gm},k}}, \sigma^+_{f^{\mathrm{ML}}_{\mathrm{gm},k}}]$, applying the usual error theory, which is based in Gaussianity of variables, because the probability distribution of each $p_{kl,\mathrm{ML}}$ is log-normal. Furthermore,

the problem is not analytic and we cannot obtain a mathematical description of the $f_{\mathrm{gm},k}$ probability distributions. We made Monte Carlo simulations to characterize the probability distribution of each $f_{\mathrm{gm},k}$. The simulations showed that the $f_{\mathrm{gm},k}$ distributions can be fit with a log-normal:

$$P_{LN}(f_{\mathrm{gm},k}|f_{\mathrm{gm},k}^{\mathrm{ML}},\sigma) = \frac{\mathrm{e}^{-(\ln f_{\mathrm{gm},k} - \ln f_{\mathrm{gm},k}^{\mathrm{ML}})^2/2\sigma^2}}{\sqrt{2\pi} f_{\mathrm{gm},k} \cdot \sigma}, \tag{21}$$

where σ is the only free parameter on the fit. Finally, the 68% confidence interval of $f_{\mathrm{gm},k}^{\mathrm{ML}}$ is given by

$$\sigma^{-}_{f_{\mathrm{gm},k}^{\mathrm{ML}}} = \mathrm{e}^{-\sigma} f_{\mathrm{gm},k}^{\mathrm{ML}}, \tag{22}$$

$$\sigma^{+}_{f_{\mathrm{gm},k}^{\mathrm{ML}}} = \mathrm{e}^{\sigma} f_{\mathrm{gm},k}^{\mathrm{ML}}. \tag{23}$$

3. SIMULATIONS WITH SYNTHETIC CATALOGS

The accuracy and reliability of the ML method can be tested using synthetic catalogs. This is an important step since ML theory warns that the estimated parameters may suffer from biases; convergence is only assured for large number of sources. The approach is to create catalogs with predefined underlying distribution parameters and compare with the estimated ML parameters. Note that the inputs of the ML method are the same whether we have a real catalog or a synthetic one. In the following paragraphs, we first explain how we created the synthetic catalogs in a general case, and later define and justify the input parameters used for the synthetic catalogs in this paper. Given the high number of variables used in the following discussion, we provide their precise definitions in Table 2.

We created the synthetic catalogs as follows: first we took n sources distributed in redshift and asymmetry space following a bidimensional distribution defined by the input probabilities $p_{kl,\mathrm{in}}$. This process yielded the $z_{\mathrm{in},i}$ and $A_{\mathrm{in},i}$ values of the n sources of our synthetic catalogs, which play the role of $z_{\mathrm{real},i}$ and $A_{\mathrm{real},i}$ in Equation 4. Next, we applied the experimental errors: following Equation 5 we obtained the simulated $z_{\mathrm{sim},i}$ values, which play the role of $z_{\mathrm{obs},i}$, as drawn from a Gaussian distribution with mean $z_{\mathrm{in},i}$ and standard deviation $\sigma_{z_{\mathrm{sim},i}}$; the latter plays the role of $\sigma_{z_{\mathrm{obs},i}}$. The value of $\sigma_{z_{\mathrm{sim},i}}$ is a positive value obtained also from a Gaussian distribution with mean $\overline{\sigma_z}$ and standard deviation σ_{σ_z}. The process was repeated following Equation 6 to obtain the simulated $A_{\mathrm{sim},i}$ and its standard deviation $\sigma_{A_{\mathrm{sim},i}}$. Finally, we applied the ML method over the synthetic catalog to obtain $p'_{kl,\mathrm{ML}}$ and $\sigma_{p'_{kl,\mathrm{ML}}}$. Summarizing, the input parameters of our simulations were the bidimensional distribution $p_{kl,\mathrm{in}}$, n, $\overline{\sigma_z}$, σ_{σ_z}, $\overline{\sigma_A}$, and σ_{σ_A}, and the output parameters were $p'_{kl,\mathrm{ML}}$ and $\sigma_{p'_{kl,\mathrm{ML}}}$.

We defined three intervals in redshift ($k = 0, 1, 2$) with $\Delta z = 0.4$ and $z \in [0, 1.2)$, and two in asymmetry ($l = 0, 1$) with $\Delta A = 0.7$ and $A \in [-0.35, 1.05)$. Distorted sources with $A > A_{\mathrm{m}} = 0.35$ (see C03 for details about the determination of this limit value) are described by $p'_{k1,\mathrm{in}}$, while normal sources by $p'_{k0,\mathrm{in}}$. We list in Table 1 the redshift and asymmetry intervals, as well as the probabilities $p_{kl,\mathrm{in}}$ and $p'_{kl,\mathrm{in}} = \ln p_{kl,\mathrm{in}}$, that define the input bidimensional distribution of our synthetic catalogs. The $p'_{kl,\mathrm{in}}$ values in Table 1 do not match any particular observational determination of these quantities, but they follow the general behavior inferred from observed galaxy merger fractions: highly asymmetric galaxies are less frequent than low-asymmetry galaxies up to $z = 1.2$ (Conselice et al. 2003; Cassata et al. 2005; Bridge et al. 2007; Kampczyk et al. 2007), so the $p'_{k1,\mathrm{in}}$ are lower than the $p'_{k0,\mathrm{in}}$. The number of highly asymmetric galaxies increases with redshift in the range $z \in [0, 1.2)$ (Conselice et al. 2003), so $p'_{k1,\mathrm{in}}$ increase with redshift. Several studies present a maximum at intermediate z in the redshift distribution of galaxies in optically selected samples (e.g., Grazian et al. 2006), so $p'_{k0,\mathrm{in}} + p'_{k1,\mathrm{in}}$ values have a maximum in the interval $z = [0.4, 0.8)$. We can check that the $p'_{kl,\mathrm{in}}$ are normalized following Equation 10. Although we preset here this particular bidimensional distribution we carried out the same study with other distributions, and the results were similar.

For convenience we express the experimental dispersions using the dimensionless variables

$$\sigma_{\mathrm{bin},z} = \frac{\overline{\sigma_z}}{\Delta z}, \tag{24}$$

$$\sigma_{\mathrm{bin},A} = \frac{\overline{\sigma_A}}{\Delta A}. \tag{25}$$

We used the same value of both variables in each simulation, that is, we used $\sigma_{\mathrm{bin}} = \sigma_{\mathrm{bin},z} = \sigma_{\mathrm{bin},A}$.

Because we fixed the values of $\Delta z = 0.4$ and $\Delta A = 0.7$, $\sigma_{\rm bin}$ unequivocally defines $\overline{\sigma_z}$ and $\overline{\sigma_A}$. It is important to notice that, when we work with observational data, the situation is the opposite: our data define $\overline{\sigma_z}$ and $\overline{\sigma_A}$, and we should choose the most appropriate values of Δz and ΔA. We made simulations for $\sigma_{\rm bin} = 0$ as a check corresponding to null experimental errors, $\sigma_{\rm bin} = 0.25$ and 0.5 as typical observational cases, and $\sigma_{\rm bin} = 1.0$ as extreme case to explore the applicability limits of the ML method. The values of σ_{σ_z} and σ_{σ_A} were a half of $\overline{\sigma_z}$ and $\overline{\sigma_A}$ respectively in all cases.

We ran models with $n = 50$, 100, and 1000 to check catalog size effects. We took these values because we expect experimental catalogs of a few hundred sources or more and we are interested in the applicability limits of the method to small samples.

In order to study how the ML parameters compare with the input parameters, we must preform several simulations and study how the parameters $p'_{kl,\rm ML}$ are distributed. Hence, for each n and $\sigma_{\rm bin}$ case we create a simulation set of $N = 1000$ independent synthetic catalogs.

The results of the simulations are shown in Figure 1, and in Tables 3, 4, and 5. Figure 1 shows $p'_{kl,\rm ML}$ for samples of $n = 1000$ sources (crosses), with error bars showing their 68% confidence intervals; for comparison, the input probabilities $p'_{kl,\rm in}$ are shown as black circles, and the $p'_{kl,\rm class}$, obtained by drawing a classical histogram (as defined below in Section 3.1), are shown as gray triangles, also for $n = 1000$ catalogs. In Figure 1, panels a, b, and c correspond to increasing values of the experimental errors, defined in Equations 24, 25 and shown in the legend; panels a, b, c may be taken to respectively describe 'good', 'typical', and 'bad' observational errors as compared to the z and A bin sizes. The top/bottom panels show p'_{kl} for the low/high-asymmetry bins. Within each panel, values for the three redshift bins are shown, as labeled on the horizontal axes. We provide the results in tabular format in Tables 3, 4, and 5, corresponding to simulations with sample sizes of $n = 50, 100$, and 1000, respectively.

3.1. Classical bidimensional distribution

Before presenting the results of the ML method, we analyze the estimation of the p'_{kl} parameters using the classical bidimensional historgram of the $z_{{\rm sim},i}$ and $A_{{\rm sim},i}$ data. We translate the histogram occupation numbers n_{kl} to probabilities $p'_{kl,\rm class}$ using

$$p'_{kl,\rm class} = \ln\left(\frac{n_{kl}}{\Delta z \Delta A \Sigma_k \Sigma_l n_{kl}}\right), \qquad (26)$$

where n_{kl} is the number of sources whit $z_{{\rm sim},i}$, $A_{{\rm sim},i}$ whitin the $[z_k, z_{k+1}) \cup [A_l, A_{l+1})$ bin. We want to study how the classical method compares with the input parameters. The distribution of the N values of $p'_{kl,\rm class}$ in one simulation set can be represented by its median $\overline{p'_{kl,\rm class}}$ and standard deviation $s_{p'_{kl,\rm class}}$. In Tables 3 - 5 we can see that the classical bidimensional distribution recovers the input probabilities in the case of null experimental errors and n large as expected. However, the shape of the input bidimensional distribution begins to deviate when $\sigma_{\rm bin}$ increases, as we can also see in Figure 1: the classical bidimensional distribution (gray triangles) is smoothed by experimental errors and does not estimate well the underlying bidimensional distribution (black circles). We study this in detail in Section 3.3.

3.2. The ML method in absence of experimental errors

We first test that the ML method, in the case of null experimental errors, recovers the input bidimensional distribution, i.e., that it reduces to the classic method. We can see in Tables 3 - 5 that the values of $\overline{p'_{kl,\rm class}}$ and the median of the N values recovered by the ML method, denoted as $\overline{p'_{kl,\rm ML}}$, are the same in all cases. This also happens with the values of $s_{p'_{kl,\rm class}}$ and the standard deviations of $\overline{p'_{kl,\rm ML}}$, denoted as $s_{p'_{kl,\rm ML}}$. This indicates that the ML method does not introduce systematic effects on the results.

3.3. The ML method with non-null experimental errors

We now examine how well the ML and classical methods recover the input probabilities $p'_{kl,\rm in}$ when non-null experimental errors are included in the synthetic catalogs. We use the $N = 1000$ source

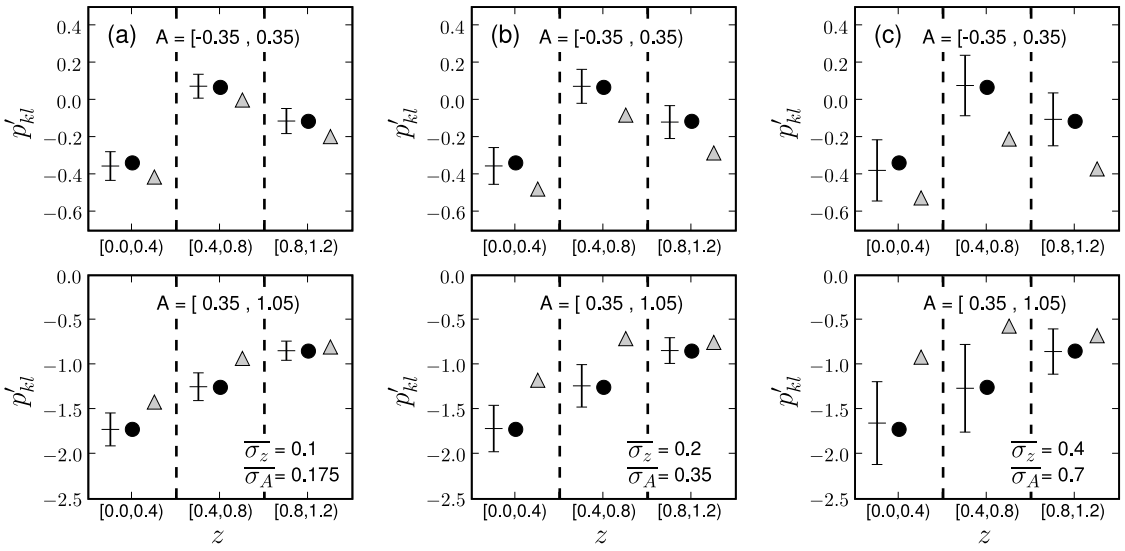

Fig. 1.— Results of run the ML method over $N = 1000$ synthetic catalogs with $n = 1000$ sources each for different experimental errors: (a) $\sigma_{\rm bin} = 0.25$, (b) $\sigma_{\rm bin} = 0.5$, and (c) $\sigma_{\rm bin} = 1$. In all figures black circles are the input bidimensional probabilities $p'_{kl,\rm in}$, gray triangles are the classical bidimensional probabilities $\overline{p'_{kl,\rm class}}$ and crosses are the ML bidimensional probabilities $\overline{p'_{kl,\rm ML}}$. The error bars are the 68% confidence intervals given by ML method, $[\overline{p'_{kl,\rm ML}} - \overline{\sigma_{p'_{kl,\rm ML}}}, \overline{p'_{kl,\rm ML}} + \overline{\sigma_{p'_{kl,\rm ML}}}]$.

catalogs as an example, which is representative of the general trends. The results are shown in Figure 1, and are tabulated in Table 5. It is clear from Figure 1 that $p'_{kl,\rm ML}$ (crosses), recover the input probabilities $p'_{kl,\rm in}$ (black circles) in all cases, including those in which the inserted errors are as large as the bin size (panels c). From Table 5 we see that the values of $p'_{kl,\rm in}$ always lay within the 68% confidence interval of the ML method, defined by $[\overline{p'_{kl,\rm ML}} - s_{p'_{kl,\rm ML}}, \overline{p'_{kl,\rm ML}} + s_{p'_{kl,\rm ML}}]$. This shows that the ML method is reliable. In contrast, the probabilities $p'_{kl,\rm class}$ derived from the classical histogram (gray triangles in Figure 1) systematically deviate from the input probabilities. Probabilities are systematically underestimated/overestimated in the low/high-asymmetry bins (upper/lower panels), due to a spill-over from the most populated bins (low asymmetries) to the least populated, high-asymmetry bins. Such deviations increase for larger experimental errors. When the errors are as large as the bin size, spill-over is so pronounced that the probabilities in the high-asymmetry sample (lower right panel) are nearly equal for the three redshift bins, and all information on the redshift variation of the galaxy merger fractions is lost.

We conclude that the ML method is an unbiased estimator of the input distribution. To put this statement in a more quantitative basis, we carry out a Student's t-test (Collins 1990, p. 232). We define our estimator as

$$T_{kl,\rm ML} = \frac{\sqrt{N}\left|p'_{kl,\rm in} - \overline{p'_{kl,\rm ML}}\right|}{s_{p'_{kl,\rm ML}}}, \qquad (27)$$

and accept that $p'_{kl,\rm in} = \overline{p'_{kl,\rm ML}}$ with a 99% of confidence when $T_{kl,\rm ML} \leq 2.6$. We define in the same way the variable $T_{kl,\rm class}$ to study the accuracy of the $p'_{kl,\rm class}$ as an estimator of the $p'_{kl,\rm in}$. We calculate the median of the $T_{kl,\rm ML}$ and $T_{kl,\rm class}$ for each simulation set, denoted as $T_{\rm ML}$ and $T_{\rm class}$ respectively, to make a comparison beetwen different n and $\sigma_{\rm bin}$.

The results are summarized in Tables 3 - 5, and in Figure 2. We can see that $T_{\rm ML}$ is below the confidence level for all n and $\sigma_{\rm bin}$: the $p'_{kl,\rm ML}$

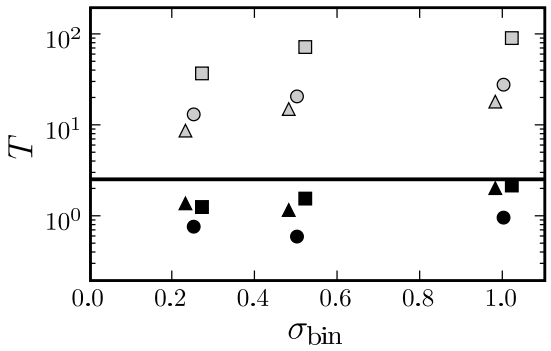

Fig. 2.— Variation of $T_{\rm ML}$ (black symbols) and $T_{\rm class}$ (gray symbols) with dimensionless experimental error size $\sigma_{\rm bin}$. Triangles are for $n = 50$, circles for $n = 100$, and squares for $n = 1000$ source catalogs. The solid line is the 99% confidence limit $T = 2.6$.

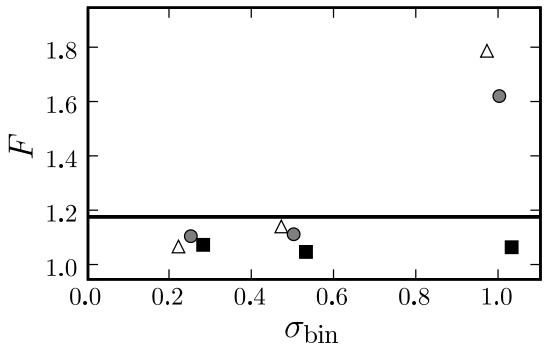

Fig. 3.— Variation of F with dimensionless experimental error size $\sigma_{\rm bin}$. Triangles are for $n = 50$, circles for $n = 100$, and squares for $n = 1000$ source catalogs. The solid line is the 99% confidence limit $F = 1.8$.

are good estimators of the $p'_{kl,\rm in}$, as wanted. In contrast, the classical method is far from the confidence condition even in the $\sigma_{\rm bin} = 0.25$ case, and $T_{\rm class}$ increases with $\sigma_{\rm bin}$. Besides, having a large n does not improve the results of classical method: the $p'_{kl,\rm class}$ values are similar for every n, but the errors are reduced when increasing n, making $T_{\rm class}$ higher. That is, having a large observational sample affected by experimental errors does not improve the estimation of $p'_{kl,\rm in}$, and the $p'_{kl,\rm class}$ errors are underestimated. This bias affects the galaxy merger fractions obtained from $p'_{kl,\rm class}$, as we can see on Section 3.7.

3.4. Study of $\sigma_{p'_{kl}}$

When we apply the ML method to an observational sample we obtain an estimation of the $p'_{kl,\rm ML}$ 68% confidence intervals, $[p'_{kl,\rm ML} - \sigma_{p'_{kl,\rm ML}}, p'_{kl,\rm ML} + \sigma_{p'_{kl,\rm ML}}]$, and we want to know if these confidence intervals are representative of the p'_{kl} probability distributions. They are representative if the median of the N values of $\sigma_{p'_{kl,\rm ML}}$, denoted as $\overline{\sigma_{p'_{kl,\rm ML}}}$, are similar to $s_{p'_{kl,\rm ML}}$. To study this issue we perform a Fisher's variance test (Collins 1990, p. 234). We define our estimator as

$$F_{kl} = \frac{\max(\overline{\sigma_{p'_{kl,\rm ML}}}, s_{p'_{kl,\rm ML}})^2}{\min(\overline{\sigma_{p'_{kl,\rm ML}}}, s_{p'_{kl,\rm ML}})^2}, \qquad (28)$$

and accept that $s_{p'_{kl,\rm ML}} = \overline{\sigma_{p'_{kl,\rm ML}}}$ with a 99% of confidence when $F_{kl} \leq 1.18$. We calculate the median of the F_{kl} for each simulation set, denoted as F, to make a comparison beetwen different n and $\sigma_{\rm bin}$. The results are summarized in Tables 3 - 5, and in Figure 3. We can see that $s_{p'_{kl,\rm ML}} = \overline{\sigma_{p'_{kl,\rm ML}}}$ for all n when $\sigma_{\rm bin} = 0.25, 0.5$. Only when $\sigma_{\rm bin} = 1.0$ and the samples are small ($n = 50, 100$) does F lie above the confidence limits.

These results imply that the ML method supplies reliable confidence intervals of $p'_{kl,\rm ML}$ with thousand sources samples or, with less sources, if the experimental errors are at most a half of the histogram bin size.

The differences between $s_{p'_{kl,\rm ML}}$ and $\overline{\sigma_{p'_{kl,\rm ML}}}$ have two origins. The main effect comes from the fact that the probability distributions of $p'_{kl,\rm ML}$ are not perfectly Gaussian, and we had assumed Gaussianity to obtain $\sigma_{p'_{kl,\rm ML}}$ analytically. We study this issue in the next section. The other effect is that we evaluated the theoretical values of $\sigma_{p'_{kl,\rm ML}}$ at $p'_{kl,\rm ML}$: the minimization method `AMOEBA` is not perfect and we may have estimated a local minimum of Equation 11 instead the absolute minimum (see Section 3.6).

3.5. Probability distributions of p'_{kl}

In the analytical estimation of the $p'_{kl,\rm ML}$ covariances we assumed that the $p'_{kl,\rm ML}$ probabil-

ity distributions are Gaussian. To check this assumption we made a histogram of the N values of $p'_{kl,\mathrm{ML}}$ to obtain the shape of the $p'_{kl,\mathrm{ML}}$ probability distribution, which we want to approximate by a Gaussian with mean $\overline{p'_{kl,\mathrm{ML}}}$ and standard deviation $s_{p'_{kl,\mathrm{ML}}}$. We tested this Gaussian approximation with a Kolmogorov-Smirnov test (Collins 1990, p. 235).

We saw that the Gaussian distribution approximation was valid for all σ_{bin} in the $n = 1000$ simulation sets. The situation of the $n = 50$ and 100 simulation sets was more complicated. For $n = 100$ the $p'_{k0,\mathrm{ML}}$ Gaussian approximation was valid for all σ_{bin}, while the $p'_{k1,\mathrm{ML}}$ started to be non Gaussian for $\sigma_{\mathrm{bin}} = 0.5$, and we could not assume Gaussianity for $\sigma_{\mathrm{bin}} = 1.0$. For $n = 50$ simulations we could not assume Gaussian approximation from $\sigma_{\mathrm{bin}} = 0.25$ to the $p'_{k1,\mathrm{ML}}$ and from $\sigma_{\mathrm{bin}} = 0.5$ to the $p'_{k0,\mathrm{ML}}$.

These results emphasize that one must check the Gaussian approximation of the $p'_{kl,\mathrm{ML}}$ probability distributions in each case. That is, when applying the ML method to an experimental catalog it is essential to make special simulations aimed at verifying the Gaussianity of the recovered probabilities.

3.6. The standard deviation of the ML method due to iterative minimization

The iterative minimization method `AMOEBA` used to obtain the minimum of Equation 11 can introduce an error in the determination of $p'_{kl,\mathrm{ML}}$ if the method converges to a local minimum. Besides, increasing the experimental errors relaxes the conditions over the absolute minimum and makes it more probable that the method converges onto one such local minimum. To study this effect and its importance, we apply the ML method $N = 100$ times over the same catalog, one per simulation set. We define the variable $s_{p'_{kl,\mathrm{iter}}}$ as the dispersion of the N values of the recovered probabilities $p'_{kl,\mathrm{ML}}$. We find that the values of $s_{p'_{kl,\mathrm{iter}}}$ depend on the tolerance and the maximum number of iterations of the minimization method. We take a 10^{-15} tolerance and 5000 iterations as optimal values: less tolerance or more iterations does not reduce $s_{p'_{kl,\mathrm{iter}}}$, but increased the computational time. All final simulations presented in this paper were made with these optimal values. We also find that $s_{p'_{kl,\mathrm{iter}}}$ increases with σ_{bin}, but is ~ 5 times smaller than $s_{p'_{kl,\mathrm{ML}}}$ in the worst experimental error case, so the standard deviations of the probabilities are slightly affected by this effect. Therefore, when applying the ML method to an experimental catalog, it is safe practice to apply it more than once, as a precaution against local solutions and iteration bias.

3.7. The galaxy merger fraction

In the previous sections we have seen that the experimental errors modify the input bidimensinal distribution, biasing the classical method estimations, whereas the ML method is able to recover the input bidimensional distribution. In this section we study the general effect and trends that the experimental errors introduce on the galaxy merger fraction determination. To obtain the galaxy merger fraction by the ML method we follow Section 2.2. First we determine the galaxy merger fraction $f^{\mathrm{ML}}_{\mathrm{gm},k}$ applying Equation 16 to the $p'_{kl,\mathrm{ML}}$ probabilities in Tables 3 - 5. Next, we perform Monte Carlo simulations with this $f^{\mathrm{ML}}_{\mathrm{gm},k}$ values and the $p'_{kl,\mathrm{ML}}$ and $\sigma_{p'_{kl,\mathrm{ML}}}$ in Tables 3 - 5 to characterize the probability distribution of $f_{\mathrm{gm},k}$, obtaining the 68% confidence interval $[\sigma^{-}_{f^{\mathrm{ML}}_{\mathrm{gm},k}}, \sigma^{+}_{f^{\mathrm{ML}}_{\mathrm{gm},k}}]$ with Equations 22 and 23.

The galaxy merger fraction by the classical method is, applying Equation 2,

$$f^{\mathrm{class}}_{\mathrm{gm},k} = \frac{2\mathrm{e}^{p'_{k1,\mathrm{class}}}}{\mathrm{e}^{p'_{k0,\mathrm{class}}} + 2\mathrm{e}^{p'_{k1,\mathrm{class}}}}, \qquad (29)$$

while its 68% confidence interval $[f^{\mathrm{class}}_{\mathrm{gm},k} - \sigma_{f^{\mathrm{class}}_{\mathrm{gm},k}}, f^{\mathrm{class}}_{\mathrm{gm},k} + \sigma_{f^{\mathrm{class}}_{\mathrm{gm},k}}]$ is obtained applying the usual error theory to Equation 29,

$$\sigma_{f^{\mathrm{class}}_{\mathrm{gm},k}} = \frac{2\mathrm{e}^{p'_{k0,\mathrm{class}}}\mathrm{e}^{p'_{k1,\mathrm{class}}}}{(\mathrm{e}^{p'_{k0,\mathrm{class}}} + 2\mathrm{e}^{p'_{k1,\mathrm{class}}})^2} \times \sqrt{s^2_{p'_{k1,\mathrm{class}}} + s^2_{p'_{k0,\mathrm{class}}}}. \qquad (30)$$

Because of the experimental error limits of the ML method which we noticed in the previous sections, we only made this study with the $n = 1000$ simulation sets. We summarize the results in Table 6, and Figure 4. We can see that the classical method gives worst estimates of the input galaxy merger fraction when the experimental errors increase. We may take as observational reference the

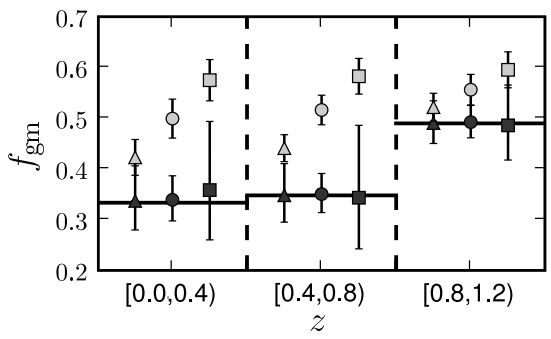

Fig. 4.— Galaxy merger fraction estimations by classical (gray symbols) and ML method (black symbols). In the two cases triangles are for $\sigma_{\rm bin} = 0.25$, circles for $\sigma_{\rm bin} = 0.5$, and squares for $\sigma_{\rm bin} = 1$. The black solid lines are the input galaxy merger fraction in each redshift bin. We can take $\sigma_{\rm bin} = 0.25$ as observational reference.

$\sigma_{\rm bin} = 0.25$ case (for example, in Conselice et al. 2003 we have $\sigma_{\rm bin} \sim 0.2$). In this case, the difference between the input and the classical estimation is ~ 0.1 on the first and second redshift intervals, which have the lower input galaxy merger fraction, and ~ 0.05 in the third interval. Furthermore, the experimental errors tend to smooth the galaxy merger fraction values. An extreme case is $\sigma_{\rm bin} = 1$, where the dependency in z has been lost. In addition, the confidence intervals are underestimated and are ~ 0.035 in every case. In contrast, the differences between the input and ML method galaxy merger fractions are ~ 0.01 in every redshift bin and experimental error case. Furthermore, the 68% confidence intervals are more realistic: in the $\sigma_{\rm bin} = 0.25, 0.5$ cases they are ~ 0.05, while in the $\sigma_{\rm bin} = 1.0$ case they increase to ~ 0.1.

Finally, we also determined the classical galaxy merger fraction in the $n = 50$ and 100 cases, and noticed that the values of $f^{\rm class}_{{\rm gm},k}$ were similar in each $\sigma_{\rm bin}$ case: having large samples does not improve the results and we must take into account the experimental errors in our analysis to avoid the bias.

4. DETERMINATION OF ANY ONE- OR BIDIMENSINAL DISTRIBUTION BY THE ML METHOD

The method outlined here may easily be applied to the unbiased determination of any bidimensional distribution in the presence of observational errors. For example, the automatic indices M_{20} and G are used in Lotz et al. (2008) to determine the galaxy merger fraction by morphological criteria. We could apply the ML method by defining the variable $MG = G + 0.14M_{20} - 0.33$ and by calling merger systems all sources with $MG > 0$. Similarly, we may apply the ML method to obtain density of sources in color-color diagrams, especially when we have some condition that separates populations, or to determine one-dimensional histogram of any observational magnitude.

For reference, we provide an outline for the application of the ML method to any one- or bidimensional experimental distribution subject to observational errors:

1. Define the observational catalog. This catalog cannot be restricted to the interval of interest, e.g., $[z_0, z_k]$, because there are sources both with $z_i < z_0$ and $z_i > z_k$ that could belong to a real bidimensional distribution bin within the range of interest due to the observational errors. In general one should include in the sample those sources with $z_i + 2\sigma_i > z_0$ and $z_i - 2\sigma_i < z_k$ to avoid incompleteness effects.

2. Apply the ML method to the observational catalog. First, define the bidimensional distribution bins taking into account the size of the observational errors. Next, minimize Equation 11 to obtain the most probable values of p'_{kl}, $p'_{kl,{\rm ML}}$. To determine their confidence intervals, calculate the Hessian matrix, Equation 13, with the observational data and the previous $p'_{kl,{\rm ML}}$ values. The diagonal elements of the inverse Hessian matrix provide $\sigma_{p'_{kl,{\rm ML}}}$. Notice that we assumed Gaussian experimental errors, Equations 5 and 6, in the development of the ML method. If you need to assume other experimental error distributions, you need to recalculate Equations 11, 14 and 15 with the new error distributions.

3. Check the results with representative synthetic catalogs. Run simulations with synthetic catalogs to test the accuracy and Gaussianity limits of the method in each particular case following the methodology of sections 3.3, 3.4 and 3.5. These synthetic catalogs should have the previous $p'_{kl,\mathrm{ML}}$ as bidimensional distribution input, that is, as $p'_{kl,\mathrm{in}}$, and similar characteristics to the experimental ones to fix the other input parameters. For example, synthetic and experimental catalogs should have same number of sources n, and $\overline{\sigma_z}$ may be given by the median of the photometric redshift errors in each redshift bin, while, for σ_{σ_z}, one may use the dispersions of these photometric redshift errors. Besides, is important to take into account special cases, e.g., the number of sources with z_{spec}, which have $\sigma_z \sim 0$, in each bin, or avoid unphysical values, e.g., negative redshifts.

4. Determine p_{kl}, Equation 16, and their confidence intervals, Equations 18 and 19, in the reliable cases.

5. CONCLUSIONS

We have presented a maximum likelihood method to recover bidimensional distributions of experimental data subject to measurement errors, and applied it to the determination of the galaxy merger fraction based on asymmetry criteria from C03.

The Gaussianity of $p'_{kl,\mathrm{ML}}$ is the strongest condition on the reliability of the method. From the results, taking into account that typical observational catalogs usually have a few hundred sources, and that the probabilities p'_{k1} would be small, we conclude that the bin of the bidimensional distribution must be at least twice the typical error in redshift in the observational catalog. Within this quality limit, the ML method can recover with accuracy and reliability the lost information due to the experimental errors. Besides, our results have realistic errors with known shapes, which the classical histograms cannot provide.

The ML method presented here may in principle be extended to as many dimensions as required by the astrophysical problem we are addressing. For instance, if we wish to determine variations in the galaxy merger fraction as a function of galaxy mass, errors in the galaxy mass determination would make objects spill over from one mass bin to the next, biasing the classical histogram approach. The ML method with an added mass axis would solve the problem. Even if we are not seeking to determine the variation of the galaxy merger fraction with mass, our parent sample unavoidably has a boundary (e.g., luminosity; mass; color), and observational errors make objects jump in and out of the sample, hence potentially modifying the shape of the distribution we are trying to determine. This extension to higher dimensions is straightforward only when the third variable is independent from the other two. In the case of a third luminosity or mass axis, this is unfortunately not the case: luminosity and mass depend on galaxy redshift, introducing covariances between the variables. Furthermore, luminosity and mass are affected by incompleteness functions, making our problem non-analytic. We leave the treatment of this problem for future work.

We dedicate this paper to the memory of our six IAC colleagues and friends who met with a fatal accident in Piedra de los Cochinos, Tenerife, in February 2007, with a special thanks to Maurizio Panniello, whose teachings of python were so important for this paper.

This work was supported by the Spanish Programa Nacional de Astronomía y Astrofísica through project number AYA2006-12955.

REFERENCES

Abraham, R. G., Tanvir, N. R., Santiago, B. X., Ellis, R. S., Glazebrook, K. & van der Bergh, S. 1996, MNRAS, 279, L47

Arzner, K., et al. 2007, A&A, 468, 501

Bell, E. F., et al. 2006, ApJ, 652, 270

Bridge, C. R., et al. 2007, ApJ, 659, 931

Bundy, K., Fukugita, M., Ellis, R. S., Kodama, T., & Conselice, C. J. 2004, ApJ, 601, L123

Cassata, P., et al. 2005, MNRAS, 357, 903

Collins, G. W. 1990, Fundamental numerical methods and data analysis, by George W. Collins, II.

Conselice, C. J., Bershady, M. A., Dickinson, M., Papovich, C. 2003, AJ, 126, 1183

Conselice, C. J. 2003, ApJS, 147, 1

—. 2006, ApJ, 638, 686

Davidson, R., & Mackinnon, J. 1993, Estimation and inference in econometrics (Ed. Oxford University Press, New York)

Davis, M., et al. 2007, ApJ, 660, L1

De Lucia, G., & Blaizot, J. 2007, MNRAS, 375, 2

De Propris, R., Liske, J., Driver, S. P., Allen, P. D., & Cross, N. J. G. 2005, AJ, 130, 1516

De Propris, R., et al. 2007, ApJ, 666, 212

Efstathiou, G. 2004, MNRAS, 348, 885

García-Dabó, C. E. 2002, Estudio estadístico de la formación estelar en el universo local (Tesis Doctoral, Universidad Complutense de Madrid)

Giavalisco, M., et al. 2004, ApJ, 600, L93

Gottlöber, S., Klypin, A., & Kravtsov, A. V. 2001, ApJ, 546, 223

Governato, F., Gardner, J. P., Stadel, J., Quinn, T., & Lake, G. 1999, AJ, 117, 1651

Grazian, A., et al. 2006, A&A, 449, 951

Kampczyk, P., et al. 2007, ApJS, 172, 329

Kartaltepe, J. S., et al. 2007, ApJS, 172, 320

Kolatt, T. S., et al. 1999, ApJ, 523, L109

Lavery, R. J., Remijan, A., Charmandaris, V., Hayes, R. D., & Ring, A. A. 2004, ApJ, 612, 679

Le Fèvre, O., et al. 2000, MNRAS, 311, 565

Lin, L., et al. 2004, ApJ, 617, L9

Lonsdale, C. J., et al. 2003, PASP, 115, 897

Lotz, J. M., et al. 2008, ApJ, 672, 177

Makarov, D., et al. 2006, AJ, 132, 2729

Marsden, J.E. & Tromba, A.J. 1996, Vector Calculus (W.H. Freeman and Company, New York)

Masjedi, M., et al. 2006, ApJ, 644, 54

Mihos, J. C. 1995, ApJ, 438, L75

Naylor, T., & Jeffries, R. D. 2006, MNRAS, 373, 1251

Nelder, J.A. & Mead, R. 1965, Computer Journal, 7(4), 308

Patton, D. R., Carlberg, R. G., Marzke, R. O., Pritchet, C. J., da Costa, L. N., & Pellegrini, P. S. 2000, ApJ, 536, 153

Patton, D. R., et al. 2002, ApJ, 565, 208

Peña, D. 2001, Fundamentos de estadística (Alianza Editorial, Madrid)

Press, W.H. 1995, Numerical Recipies in C, second edition (Cambridge University Press, New York)

Scoville, N., et al. 2007, ApJS, 172, 1

Sheth, R. K. 2007, MNRAS, 378, 709

Springel, V., Di Matteo, T., & Hernquist, L. 2005, ApJ, 620, L79

Wolf, C., Meisenheimer, K., Rix, H. -W., Borch, A., Dye, S., & Kleinheinrich, M. 2003, A&A, 401, 73

This 2-column preprint was prepared with the AAS LaTeX macros v5.2.

TABLE 1

INPUT BIDIMENSIONAL DISTRIBUTION USED FOR THE SYNTHETIC CATALOGS

k	l	$p_{kl,\mathrm{in}}$	$p'_{kl,\mathrm{in}}$	$[z_k, z_{k+1})$	$[A_l, A_{l+1})$
0	0	0.71428	-0.33647	[0, 0.4)	[-0.35, 0.35)
1	0	1.07143	0.06899	[0.4, 0.8)	[-0.35, 0.35)
2	0	0.89286	-0.11333	[0.8, 1.2)	[-0.35, 0.35)
0	1	0.17857	-1.72277	[0, 0.4)	[0.35, 1.05)
1	1	0.28571	-1.25276	[0.4, 0.8)	[0.35, 1.05)
2	1	0.42857	-0.84730	[0.8, 1.2)	[0.35, 1.05)

NOTE.—Variable definitions:
k: index that scans the redshift bins.
l: index that scans the asymmetry bins.
$p_{kl,\mathrm{in}}$: probability that a source has redshift in bin k and asymmetry in bin l.
$p'_{kl,\mathrm{in}}$: logarithm of $p_{kl,\mathrm{in}}$.
$[z_k, z_{k+1})$: redshift bin k.
$[A_l, A_{l+1})$: asymmetry bin l.

TABLE 2

VARIABLE DEFINITIONS FOR THE SIMULATIONS

Variable	Definition
	Input Variables
n	Number of total sources in a synthetic catalog.
n_{kl}	Number of sources in $[z_k, z_{k+1}) \cup [A_l, A_{l+1})$ bin.
Δz	Redshift bin size.
ΔA	Asymmetry bin size.
N	Number of synthetic catalogs in each simulation set.
$p'_{kl,\mathrm{in}}$	Logarithmic probabilities of the input bidimensional distribution of the synthetic catalogs
$\overline{\sigma_z}$	Median experimental errors in redshift of the synthetic catalog sources.
σ_{σ_z}	Dispersion on σ_z of the synthetic catalog sources.
$\overline{\sigma_A}$	Median experimental errors in asymmetry of the synthetic catalog sources.
σ_{σ_A}	Dispersion on σ_A of the synthetic catalog sources.
σ_{bin}	$\frac{\overline{\sigma_z}}{\Delta z} = \frac{\overline{\sigma_A}}{\Delta A}$. Dimensionless experimental error size.
	Output Variables
$p'_{kl,\mathrm{class}}$	Classical logarithmic probabilities of the classical bidimensional distribution.
$\overline{p'_{kl,\mathrm{class}}}$	Median of the N values of $p'_{kl,\mathrm{class}}$ in one simulation set.
$s_{p'_{kl,\mathrm{class}}}$	Standard deviation of the N values of $p'_{kl,\mathrm{class}}$ in one simulation set.
$p'_{kl,\mathrm{ML}}$	Logarithmic probabilities of the bidimensional distribution recovered by the ML method.
$\sigma_{p'_{kl,\mathrm{ML}}}$	The 68% confidence interval of $p'_{kl,\mathrm{ML}}$ given by the ML method, $[p'_{kl,\mathrm{ML}} - \sigma_{p'_{kl,\mathrm{ML}}}, p'_{kl,\mathrm{ML}} + \sigma_{p'_{kl,\mathrm{ML}}}]$.
$\overline{p'_{kl,\mathrm{ML}}}$	Median of the N values of $p'_{kl,\mathrm{ML}}$ in one simulation set.
$s_{p'_{kl,\mathrm{ML}}}$	Standard deviation of the N values of $p'_{kl,\mathrm{ML}}$ in one simulation set.
$\overline{\sigma_{p'_{kl,\mathrm{ML}}}}$	Median of the N values of $\sigma_{p'_{kl,\mathrm{ML}}}$ in one simulation set.
	Quality Variables
$T_{kl,\mathrm{ML}}$	$\frac{\sqrt{N}\left\lvert p'_{kl,\mathrm{in}} - \overline{p'_{kl,\mathrm{ML}}}\right\rvert}{s_{p'_{kl,\mathrm{ML}}}}$. Accepted that $p'_{kl,\mathrm{in}} = \overline{p'_{kl,\mathrm{ML}}}$ when $T_{kl,\mathrm{ML}} \leq 2.6$.
F_{kl}	$\frac{\max(\overline{\sigma_{p'_{kl,\mathrm{ML}}}}, s_{p'_{kl,\mathrm{ML}}})^2}{\min(\overline{\sigma_{p'_{kl,\mathrm{ML}}}}, s_{p'_{kl,\mathrm{ML}}})^2}$. Accepted that $s_{p'_{kl,\mathrm{ML}}} = \overline{\sigma_{p'_{kl,\mathrm{ML}}}}$ when $F_{kl} \leq 1.18$.
$s_{p'_{kl,\mathrm{iter}}}$	Standard deviation of the ML method due to iterative minimization process.

TABLE 3

RESULTS OF ML METHOD OVER N = 1000 SYNTHETIC CATALOGS WITH N = 50 SOURCES

p'_{kl}	$p'_{kl,\mathrm{in}}$	$\overline{p'_{kl,\mathrm{ML}}}$	$T_{kl,\mathrm{ML}}$	$s_{p'_{kl,\mathrm{ML}}}$	$\overline{\sigma_{p'_{kl,\mathrm{ML}}}}$	F_{kl}	$\overline{p'_{kl,\mathrm{class}}}$	$s_{p'_{kl,\mathrm{class}}}$	$T_{kl,\mathrm{class}}$
		$\overline{\sigma_z}=0$	$\sigma_{\sigma_z}=0$		$\overline{\sigma_A}=0$	$\sigma_{\sigma_A}=0$	$\sigma_{\mathrm{bin}}=0$		
p'_{00}	-0.33647	-0.31627	$\cdots$	0.30034	0.28284	$\cdots$	-0.31627	0.30034	$\cdots$
p'_{10}	0.06899	0.08920	$\cdots$	0.19871	0.22361	$\cdots$	0.08920	0.19871	$\cdots$
p'_{20}	-0.11333	-0.13395	$\cdots$	0.30034	0.25166	$\cdots$	-0.13395	0.30034	$\cdots$
p'_{01}	-1.72277	-1.92571	$\cdots$	0.81379	0.69282	$\cdots$	-1.92571	0.81379	$\cdots$
p'_{11}	-1.25276	-1.23256	$\cdots$	0.37839	0.47958	$\cdots$	-1.23256	0.37839	$\cdots$
p'_{21}	-0.84730	-0.82710	$\cdots$	0.51344	0.38297	$\cdots$	-0.82710	0.51344	$\cdots$
		$\overline{\sigma_z}=0.1$	$\sigma_{\sigma_z}=0.05$		$\overline{\sigma_A}=0.175$	$\sigma_{\sigma_A}=0.0875$	$\sigma_{\mathrm{bin}}=0.25$		
p'_{00}	-0.33647	-0.33337	0.29	0.33981	0.34855	1.052	-0.39768	0.34393	5.63
p'_{10}	0.06899	0.08583	1.82	0.29197	0.29191	1.001	0.00779	0.24453	7.91
p'_{20}	-0.11333	-0.11093	0.24	0.31124	0.30611	1.034	-0.20757	0.30034	9.92
p'_{01}	-1.72277	-1.78200	1.82	1.02762	0.87331	1.385	-1.40150	0.53047	19.15
p'_{11}	-1.25276	-1.32243	2.39	0.92324	0.74240	1.546	-0.93512	0.45987	21.84
p'_{21}	-0.84730	-0.83122	1.01	0.50531	0.48405	1.090	-0.82005	0.39750	2.17
		$\overline{\sigma_z}=0.2$	$\sigma_{\sigma_z}=0.1$		$\overline{\sigma_A}=0.35$	$\sigma_{\sigma_A}=0.175$	$\sigma_{\mathrm{bin}}=0.5$		
p'_{00}	-0.33647	-0.34488	0.59	0.45113	0.47976	1.131	-0.49339	0.36822	13.48
p'_{10}	0.06899	0.08788	1.37	0.43485	0.42913	1.027	-0.07188	0.29763	14.97
p'_{20}	-0.11333	-0.07862	2.66	0.41200	0.39237	1.102	-0.29420	0.36126	15.83
p'_{01}	-1.72277	-1.88189	1.05	4.79089	1.45853	10.789	-1.17049	0.55578	31.42
p'_{11}	-1.25276	-1.27478	0.31	2.25784	1.21780	3.437	-0.72192	0.44502	37.72
p'_{21}	-0.84730	-0.87929	1.34	0.75458	0.70106	1.158	-0.71523	0.46986	8.89
		$\overline{\sigma_z}=0.4$	$\sigma_{\sigma_z}=0.2$		$\overline{\sigma_A}=0.7$	$\sigma_{\sigma_A}=0.35$	$\sigma_{\mathrm{bin}}=1.0$		
p'_{00}	-0.33647	-0.31312	1.21	0.60964	0.88500	2.107	-0.53435	0.49716	12.59
p'_{10}	0.06899	0.15383	3.79	0.70757	0.85943	1.475	-0.19260	0.39926	20.72
p'_{20}	-0.11333	-0.01272	4.91	0.64776	0.69944	1.166	-0.34675	0.44899	16.44
p'_{01}	-1.72277	-2.20397	1.61	9.46236	5.43006	3.037	-0.93981	0.61357	40.35
p'_{11}	-1.25276	-1.78130	2.55	6.54863	5.83082	1.261	-0.58564	0.55583	37.95
p'_{21}	-0.84730	-0.89694	0.70	2.25333	1.44775	2.422	-0.68095	0.54790	9.60

TABLE 4

RESULTS OF ML METHOD OVER N = 1000 SYNTHETIC CATALOGS WITH N = 100 SOURCES

p'_{kl}	$p'_{kl,\mathrm{in}}$	$\overline{p'_{kl,\mathrm{ML}}}$	$T_{kl,\mathrm{ML}}$	$s_{p'_{kl,\mathrm{ML}}}$	$\overline{\sigma_{p'_{kl,\mathrm{ML}}}}$	F_{kl}	$\overline{p'_{kl,\mathrm{class}}}$	$s_{p'_{kl,\mathrm{class}}}$	$T_{kl,\mathrm{class}}$
		$\overline{\sigma_z}=0$	$\sigma_{\sigma_z}=0$		$\overline{\sigma_A}=0$	$\sigma_{\sigma_A}=0$	$\sigma_{\mathrm{bin}}=0$		
p'_{00}	-0.33647	-0.33647	$\cdots$	0.22391	0.20000	$\cdots$	-0.33647	0.22391	$\cdots$
p'_{10}	0.06899	0.06899	$\cdots$	0.14864	0.15275	$\cdots$	0.06899	0.14864	$\cdots$
p'_{20}	-0.11333	-0.11333	$\cdots$	0.17864	0.17321	$\cdots$	-0.11333	0.17864	$\cdots$
p'_{01}	-1.72277	-1.72277	$\cdots$	0.62763	0.43589	$\cdots$	-1.72277	0.62763	$\cdots$
p'_{11}	-1.25276	-1.25276	$\cdots$	0.37839	0.33912	$\cdots$	-1.25276	0.37839	$\cdots$
p'_{21}	-0.84730	-0.84730	$\cdots$	0.24924	0.27080	$\cdots$	-0.84730	0.24924	$\cdots$
		$\overline{\sigma_z}=0.1$	$\sigma_{\sigma_z}=0.05$		$\overline{\sigma_A}=0.175$	$\sigma_{\sigma_A}=0.0875$	$\sigma_{\mathrm{bin}}=0.25$		
p'_{00}	-0.33647	-0.35726	2.77	0.23715	0.24543	1.071	-0.42902	0.22794	12.84
p'_{10}	0.06899	0.08249	2.02	0.21156	0.20096	1.108	0.01114	0.18577	9.85
p'_{20}	-0.11333	-0.11930	0.81	0.23380	0.21587	1.173	-0.20244	0.20019	14.08
p'_{01}	-1.72277	-1.72171	0.05	0.61932	0.57412	1.164	-1.40984	0.41453	23.87
p'_{11}	-1.25276	-1.24035	0.76	0.51661	0.49041	1.110	-0.92149	0.28759	36.42
p'_{21}	-0.84730	-0.84335	0.37	0.33559	0.34188	1.038	-0.79798	0.28500	5.47
		$\overline{\sigma_z}=0.2$	$\sigma_{\sigma_z}=0.1$		$\overline{\sigma_A}=0.35$	$\sigma_{\sigma_A}=0.175$	$\sigma_{\mathrm{bin}}=0.5$		
p'_{00}	-0.33647	-0.34332	0.74	0.28963	0.31999	1.221	-0.47807	0.26072	17.17
p'_{10}	0.06899	0.07331	0.47	0.28921	0.29111	1.013	-0.07671	0.20306	22.69
p'_{20}	-0.11333	-0.11022	0.35	0.27894	0.27998	1.007	-0.26198	0.23834	19.72
p'_{01}	-1.72277	-1.69378	1.03	0.89067	0.81972	1.181	-1.17948	0.42601	40.33
p'_{11}	-1.25276	-1.26124	0.30	0.88586	0.77025	1.323	-0.72455	0.31834	52.47
p'_{21}	-0.84730	-0.85992	0.88	0.45226	0.46377	1.051	-0.76835	0.32399	7.71
		$\overline{\sigma_z}=0.4$	$\sigma_{\sigma_z}=0.2$		$\overline{\sigma_A}=0.7$	$\sigma_{\sigma_A}=0.35$	$\sigma_{\mathrm{bin}}=1.0$		
p'_{00}	-0.33647	-0.34227	0.40	0.45903	0.57113	1.548	-0.52269	0.33388	17.64
p'_{10}	0.06899	0.10014	2.25	0.43673	0.56965	1.701	-0.18589	0.26832	30.04
p'_{20}	-0.11333	-0.08903	1.63	0.46981	0.47632	1.028	-0.39434	0.32866	27.04
p'_{01}	-1.72277	-1.82510	0.65	4.99004	2.17650	5.256	-0.91894	0.40459	62.83
p'_{11}	-1.25276	-1.38142	1.32	3.07524	2.24256	1.880	-0.58439	0.34776	60.78
p'_{21}	-0.84730	-0.84699	0.01	0.85518	0.87014	1.035	-0.65849	0.35630	16.76

TABLE 5

RESULTS OF ML METHOD OVER N = 1000 SYNTHETIC CATALOGS WITH n = 1000 SOURCES

p'_{kl}	$p'_{kl,\mathrm{in}}$	$\overline{p'_{kl,\mathrm{ML}}}$	$T_{kl,\mathrm{ML}}$	$s_{p'_{kl,\mathrm{ML}}}$	$\overline{\sigma_{p'_{kl,\mathrm{ML}}}}$	F_{kl}	$\overline{p'_{kl,\mathrm{class}}}$	$s_{p'_{kl,\mathrm{class}}}$	$T_{kl,\mathrm{class}}$
		$\overline{\sigma_z}=0$	$\sigma_{\sigma_z}=0$		$\overline{\sigma_A}=0$	$\sigma_{\sigma_A}=0$	$\sigma_{\mathrm{bin}}=0$		
p'_{00}	-0.33647	-0.33647	$\cdots$	0.06137	0.06325	$\cdots$	-0.33647	0.06137	$\cdots$
p'_{10}	0.06899	0.06899	$\cdots$	0.04940	0.04830	$\cdots$	0.06899	0.04940	$\cdots$
p'_{20}	-0.11333	-0.11333	$\cdots$	0.05336	0.05477	$\cdots$	-0.11333	0.05336	$\cdots$
p'_{01}	-1.72277	-1.72277	$\cdots$	0.14864	0.13784	$\cdots$	-1.72277	0.14864	$\cdots$
p'_{11}	-1.25276	-1.25276	$\cdots$	0.11132	0.10724	$\cdots$	-1.25276	0.11132	$\cdots$
p'_{21}	-0.84730	-0.84730	$\cdots$	0.08066	0.08563	$\cdots$	-0.84730	0.08066	$\cdots$
		$\overline{\sigma_z}=0.1$	$\sigma_{\sigma_z}=0.05$		$\overline{\sigma_A}=0.175$	$\sigma_{\sigma_A}=0.0875$	$\sigma_{\mathrm{bin}}=0.25$		
p'_{00}	-0.33647	-0.35327	6.99	0.07596	0.07700	1.028	-0.41246	0.07170	33.51
p'_{10}	0.06899	0.07525	3.19	0.06204	0.06414	1.139	0.00102	0.05671	37.90
p'_{20}	-0.11333	-0.11170	0.73	0.07046	0.06730	1.094	-0.19532	0.06814	38.05
p'_{01}	-1.72277	-1.72333	0.10	0.18378	0.18398	1.005	-1.41729	0.12568	76.86
p'_{11}	-1.25276	-1.24578	1.35	0.16351	0.15516	1.138	-0.92943	0.09163	111.59
p'_{21}	-0.84730	-0.84306	1.23	0.10901	0.10726	1.060	-0.80147	0.08635	16.78
		$\overline{\sigma_z}=0.2$	$\sigma_{\sigma_z}=0.1$		$\overline{\sigma_A}=0.35$	$\sigma_{\sigma_A}=0.175$	$\sigma_{\mathrm{bin}}=0.5$		
p'_{00}	-0.33647	-0.35273	5.28	0.09732	0.09866	1.027	-0.47684	0.08066	55.03
p'_{10}	0.06899	0.07463	1.83	0.09727	0.09114	1.069	-0.07992	0.06535	72.06
p'_{20}	-0.11333	-0.11731	1.36	0.09237	0.08829	1.096	-0.28307	0.07064	75.99
p'_{01}	-1.72277	-1.71352	1.12	0.26019	0.25957	1.002	-1.17318	0.13081	132.86
p'_{11}	-1.25276	-1.23600	2.09	0.25323	0.23734	1.110	-0.70824	0.09678	177.92
p'_{21}	-0.84730	-0.84142	1.25	0.14808	0.14385	1.033	-0.75047	0.10000	30.62
		$\overline{\sigma_z}=0.4$	$\sigma_{\sigma_z}=0.2$		$\overline{\sigma_A}=0.7$	$\sigma_{\sigma_A}=0.35$	$\sigma_{\mathrm{bin}}=1.0$		
p'_{00}	-0.33647	-0.37667	8.27	0.15372	0.16437	1.143	-0.52454	0.10316	57.65
p'_{10}	0.06899	0.07901	1.95	0.16218	0.16229	1.001	-0.20842	0.08748	100.28
p'_{20}	-0.11333	-0.10270	2.45	0.13490	0.14196	1.107	-0.36879	0.09448	85.51
p'_{01}	-1.72277	-1.65260	5.03	0.44100	0.46316	1.103	-0.91542	0.13080	195.18
p'_{11}	-1.25276	-1.26323	0.67	0.49175	0.49023	1.006	-0.56759	0.11510	188.24
p'_{21}	-0.84730	-0.85242	0.65	0.24892	0.25306	1.033	-0.67546	0.11284	48.15

TABLE 6

REAL, ML METHOD, AND CLASSIC GALAXY MERGER FRACTION

$[z_k, z_{k+1})$	$f_{\mathrm{gm},k}^{\mathrm{in}}$	$f_{\mathrm{gm},k}^{\mathrm{ML}}$			$f_{\mathrm{gm},k}^{\mathrm{class}}$		
		$\sigma_{\mathrm{bin}} = 0.25$	$\sigma_{\mathrm{bin}} = 0.5$	$\sigma_{\mathrm{bin}} = 1.0$	$\sigma_{\mathrm{bin}} = 0.25$	$\sigma_{\mathrm{bin}} = 0.5$	$\sigma_{\mathrm{bin}} = 1.0$
$[0, 0.4)$	0.3333	$0.337^{+0.069}_{-0.057}$	$0.339^{+0.048}_{-0.042}$	$0.358^{+0.134}_{-0.098}$	0.423 ± 0.035	0.499 ± 0.038	0.575 ± 0.040
$[0.4, 0.8)$	0.3478	$0.348^{+0.063}_{-0.053}$	$0.350^{+0.040}_{-0.036}$	$0.343^{+0.140}_{-0.099}$	0.441 ± 0.027	0.516 ± 0.029	0.583 ± 0.035
$[0.8, 1.2)$	0.4897	$0.490^{+0.044}_{-0.040}$	$0.492^{+0.032}_{-0.030}$	$0.486^{+0.079}_{-0.068}$	0.522 ± 0.027	0.556 ± 0.030	0.595 ± 0.035